AD-A045 106 PURDUE UNIV LAFAYETTE IND DEPT OF STATISTICS F/G 12/1
ON SELECTION RULES FOR FINITE MIXTURES OF DISTRIBUTIONS, (U)
AUG 77 S S GUPTA, W HUANG N00014-75-C-0455
UNCLASSIFIED MIMEOGRAPH SER-504 NL

1 OF 1
AD
A045106

END
DATE
FILMED
11 - 77
DDC

(1)

PURDUE UNIVERSITY

DEPARTMENT OF STATISTICS

DIVISION OF MATHEMATICAL SCIENCES

6

On Selection Rules for Finite Mixtures of Distributions*

by

10 Shanti S. Gupta, Purdue University

and

Wen-Tao Huang, Academica Sinica, Taipei

9 Technical rept.,

Department of Statistics
Division of Mathematical Sciences
14 Mimeograph Series 504
11 August 1977

12 26 p.

*This research was supported by the Office of Naval Research Contract 15 N00014-75-C-0455 at Purdue University.

291 730

On Selection Rules for Finite Mixtures of Distributions*

by

Shanti S. Gupta, Purdue University

and

Wen-Tao Huang, Academia Sinica, Taipei

1. Introduction and Summary

In some industrial processes, ordinarily observations arise from some distribution function $F_1(x;\theta)$, say, but occasionally, the process yields "outliers" which may follow some other distribution $F_2(x;\varphi)$. Accordingly, if outliers carry no obvious label, then the process produces observations according to a mixture distribution $\alpha F_1(x;\theta) + (1-\alpha)F_2(x;\varphi)$ for some proportion α, $0 < \alpha < 1$. Also, in marine biology, one may be interested in studying certain characteristics of a fish. For this purpose, samples of fish are taken and the desired trait is [illegible]ed for each fish. Since many characteristics vary according to the age of fish, the trait has a distinct distribution for each age group and the population has a mixture of distributions. On the other hand, mixtures of distributions occur in the compound decision problems as proposed by Robbins [8], in which mixing distributions correspond to some a priori distributions.

It happens that in many cases, an experimenter is faced with a problem of choosing one or more "desirable" processes (treatments) from among k given processes (treatments) which produce observations according to some

---

*This research was supported by the Office of Naval Research contract N00014-75-C-0455 at Purdue University. Reproduction in whole or in part is permitted for any purpose of the United States Government.

mixture distributions. For his special purpose, the experimenter may need one or more processes which are associated with the largest (smallest) proportion in the mixtures of distributions. For instance, he may need the process which has the least proportion of occurence of outliers.

The problem for the estimation of these proportions of mixtures is not easy. For example, when $F_1(x;\theta)$ and $F_2(x;\varphi)$ both are normal with a common variance σ^2 and, with means μ_1 and μ_2, respectively, if μ_1 and μ_2 are not well separated, i.e. when $d \equiv |\mu_1 - \mu_2|/\sigma$ is small, it is almost impossible to classify the observations from the mixture distribution into two groups. To be more precise, let $I(\alpha; F_1,F_2)$ denote the Fisher information for the estimation of α. Hill [5] pointed out that for d small, $I(\alpha;F_1,F_2) \approx d^2$. Therefore, if d is in (1/8,1/4), then for the maximum likelihood estimate for α with standard deviation 0.1, the sample size needed is large, as big as 6400. However, so far the classical efficient method for the estimation of α is the maximum likelihood estimate. The usual moment estimate is inefficient. However, when the number of components of mixture increases, situations become more complicated even for the maximum likelihood estimates. This suggests that another approach should be considered. The so-called minimum distance method of Wolfowitz [12] seems reasonable. Large sample properties like consistency can be shown to hold. If the distance between two distribution functions is properly chosen, some other optimal properties may hold. And, if the rate of convergence is fair, then this approach should be right. In the problems of selection and ranking, some statistics are necessary such that based on these quantities, the criterion of priority of selection can be constructed. Though these statistics may not necessarily be good estimates

for the unknown parameters which are under consideration, most of them may do well for the selection problem. Accordingly, the minimum distance method seems natural to be one of the approaches to follow for the problems of selecting the largest (or smallest) proportion of certain component of mixture. The so-called least squares method will be applied in this paper for the selection problem.

In section 2 some notation are defined and the problem is formulated. A class of consistent selection procedures are defined in section 3 and some asymptotic optimal properties are shown.

2. Notation and Formulation of the Problem

The problem of identifiability should be mentioned, since the selection problem for the proportion of mixtures is related to the identifiability problem. This can be simply illustrated by an example. Let $B(n;p)$ denote a binomial distribution with success probability p, then, it can be found some α_1, α_2, β_1, β_2 and some p_1, p_2 and p_3 such that $\alpha_1 \neq \beta_1$ and $\alpha_1 B(n;p_1) + \alpha_2 B(n;p_2) + (1-\alpha_1-\alpha_2)B(n;p_3)$ and $\beta_1 B(n;p_1) + \beta_2 B(n;p_2) + (1-\beta_1-\beta_2)B(n;p_3)$ represent the same mixture distribution if $n < 5$. In this example, it is impossible to identify and select. Necessary and sufficient conditions for identifiability of finite mixtures can be found in [11] and [13].

Let $\mathfrak{F}$ denote the family of distributions such that the associated convex hull of $\mathfrak{F}$ is identifiable. Many well-known families of distributions are included in $\mathfrak{F}$. For example (see [11], [13]), $\mathfrak{F}$ can be family of p-variate normal distributions, product of n exponential distributions, binomial distributions with different integral parameters, translation

parameter family induced by a certain univariate cdf, union of the families of product of n exponential distributions and the p-variate normal distribution etc.

For convenience, for some prefixed integer m, we define

(2.1) $$\langle 0,1\rangle^m = \{(\alpha_1,\alpha_2,\ldots,\alpha_m): \alpha_i > 0, \sum_1^m \alpha_i = 1\} \quad (m \geq 2).$$

Let λ be a real-valued continuous function on $\langle 0,1\rangle^m$. Let the functions $F_1(x;\theta_1), F_2(x;\theta_2),\ldots,F_m(x;\theta_m)$ be in $\mathcal{F}$, where θ_i may be a parameter vector and $F_i(x;\theta_i)$ and $F_j(x;\theta_j)$ may have different parametric form, for instance, $F_i(x;\theta_i)$ may be a normal distribution with location-scale parameter (μ_i,σ_i^2) and $F_j(x;\theta_j)$ may be an exponential distribution with location-scale parameter (α_j,β_j). For convenience, we denote

(2.2) $$\underset{\sim}{F} = (F_1(x;\theta_1),\ldots,F_m(x;\theta_m))$$

and

(2.3) $$\underset{\sim}{\alpha}_i = (\alpha_{i1},\alpha_{i2},\ldots,\alpha_{im}).$$

A finite mixture distribution with m component is defined to be the inner product of certain $\underset{\sim}{\alpha} \in \langle 0,1\rangle^m$ and $\underset{\sim}{F}$, i.e.

(2.4) $$G(x;\underset{\sim}{\alpha}) = \underset{\sim}{\alpha} \cdot \underset{\sim}{F}$$

$$= \sum_{i=1}^m \alpha_i F_i(x;\theta_i)$$

Let $\pi_1, \pi_2,\ldots,\pi_k$ be k populations such that π_i has cdf $G(x;\underset{\sim}{\alpha}_i)$ (defined by (2.4)) for some unknown parameter $\underset{\sim}{\alpha}_i \in \langle 0,1\rangle^m$. Let $X_{i1},X_{i2},\ldots,X_{im}$ be n

random observations from π_i, $i=1,2,\ldots,k$. Let $G_{in}(x)$ denote the associated empirical distribution function. Let $\lambda_{[1]}(\underset{\sim}{\alpha}) \leq \lambda_{[2]}(\underset{\sim}{\alpha}) \leq \ldots \leq \lambda_{[k]}(\underset{\sim}{\alpha})$ denote the order values of $\lambda(\underset{\sim}{\alpha}_1), \lambda(\underset{\sim}{\alpha}_2), \ldots, \lambda(\underset{\sim}{\alpha}_k)$.

Based on n independent observations from each population, we are interested in selecting t $(1 \leq t \leq k - 1)$ populations, say, $\pi_{r_1}, \pi_{r_2}, \ldots, \pi_{r_t}$ such that $\lambda(\underset{\sim}{\alpha}_{r_1}), \lambda(\underset{\sim}{\alpha}_{r_2}), \ldots, \lambda(\underset{\sim}{\alpha}_{r_t})$ are the t largest. We call these populations the t best.

We approach the problem by the indifference zone formulation. For convenience, we introduce the following notation.
For given Δ, we define

$$(2.5)\quad \Omega(\lambda;\Delta) = \{(\underset{\sim}{\alpha}_1, \underset{\sim}{\alpha}_2, \ldots, \underset{\sim}{\alpha}_k) : \underset{\sim}{\alpha}_i \in \langle 0,1\rangle^m,\ \lambda_{[k-t+1]}(\underset{\sim}{\alpha}) \geq \lambda_{[k-t]}(\underset{\sim}{\alpha}) + \Delta\}.$$

For specified $\underset{\sim}{F}$ and λ, we consider our problem on the configuration $\Omega(\lambda;\Delta)$ for given Δ for the indifference zone approach. We also define

$$(2.6)\qquad \Omega = \langle 0,1\rangle^m \times \langle 0,1\rangle^m \times \ldots \times \langle 0,1\rangle^m. \quad (k \text{ copies})$$

Finally, we define, for given p, $0 \leq p \leq 1$

$$(2.7)\qquad S(\underset{\sim}{\alpha};H) = \int_{-\infty}^{\infty} (\underset{\sim}{\alpha}\cdot F - G_n(x))^2 dH(x)$$

where $\underset{\sim}{\alpha} \cdot \underset{\sim}{F}$ is a mixture distribution for $\underset{\sim}{\alpha} \in \langle 0,1\rangle^m$ and $G_m(x)$ is the empirical distribution associated with some $\underset{\sim}{\alpha}_0 \cdot \underset{\sim}{F}$ for unknown $\underset{\sim}{\alpha}_0 \in \langle 0,1\rangle^m$. And $H(x)$ is a cdf. Hence, $S(\underset{\sim}{\alpha};H)$ is a function on $\underset{\sim}{\alpha} \in \langle 0,1\rangle^m$.

3. A class of consistent selection procedures

In this section, we consider the cases when F are continuous and discrete. In each case, we assume the component $F_i(x_i;\theta_i)$ of F are completely known.

(A) Continuous case

We assume the parametric form of each component $F_i(x;\theta_i)$ of F is continuous in x for each θ_i and continuous in θ_i for each x.

For given n observations from a population with cdf $G(x;\underset{\sim}{\alpha}_0) = \underset{\sim}{\alpha}_0 \cdot \underset{\sim}{F}$ for some unknown $\underset{\sim}{\alpha}_0$ and a given cdf $H(x)$, a vector $\hat{\underset{\sim}{\alpha}} \in \langle 0,1\rangle^m$ at which $S(\underset{\sim}{\alpha};H)$ attains its infimum seems a "good" estimate for the real $\underset{\sim}{\alpha}_0$ in the sense of least squares method. It is to be noted that $\hat{\underset{\sim}{\alpha}}$ is a statistic of n observations and also is a function of $\underset{\sim}{F}$ and H. A good choice in some sense for the weight function $H(x)$ is not easy. Bartlett and Macdonald [1] study some special case of $m = 2$. For $m \geq 3$, the situation is complicated.

Choi and Balgren [3] consider the case $H(x) = G_{in}(x)$ and obtain some optimal properties like consistency and asymptotical normality. However, for the case of small samples, Macdonald [7] points out that, using $H(x) = \underset{\sim}{\alpha} \cdot \underset{\sim}{F}$, some Monte Carlo results show some improvement of the Choi and Bulgren's result. And, as a matter of fact, for $H(x) = \underset{\sim}{\alpha} \cdot \underset{\sim}{F}$, $S(\underset{\sim}{\alpha};H)$ is the von Mises statistic for the goodness-of-fit. Let U_1 and U_2 denote two random observations from the population with cdf $F(x)$ and V_1 and V_2 denote the random observations from a population with cdf $G(x)$. It is known that $\Delta(F,G) \equiv P_r\{ U_1 \vee U_2 < V_1 \wedge V_2 \text{ or } V_1 \vee V_2 < U_1 \wedge U_2\} = 1/3 + 1/2 \int_{-\infty}^{\infty} (F(x) - G(x))^2 \, d(\frac{F(x)+G(x)}{2})$ where $a \vee b = \max(a,b)$, $a \wedge b = \min(a,b)$ (Lehmann [6]). Note that $\Delta(F,G) = 0$ if, and only if $F \equiv G$. Roughly speaking, taking $F(x)$ to be $\underset{\sim}{\alpha} \cdot \underset{\sim}{F}$ and $G(x)$ to be $G_m(x)$, it is significant to consider $H(x) = \frac{1}{2}(\underset{\sim}{\alpha} \cdot \underset{\sim}{F} + G_m(x))$ for our case. Accordingly, in general, we consider the case $H(x) = p\,\underset{\sim}{\alpha} \cdot \underset{\sim}{F} + (1-p)G_m(x)$ for $0 \leq p \leq 1$. Note that

$p = 0$ yields the Choi and Bulgren's case and for $p = 1$, we get the Macdonald's case. For our notational convenience, henceforth, we define

$$(3.1) \qquad S_i(\underset{\sim}{\alpha};p) = \int_{-\infty}^{\infty} (\underset{\sim}{\alpha}\cdot\underset{\sim}{F} - G_{in}(x))^2 d(p\,\underset{\sim}{\alpha}\cdot\underset{\sim}{F} + (1-p)G_{in}(x))$$

which is obtained by taking $H(x) = p\,\underset{\sim}{\alpha}\cdot\underset{\sim}{F} + (1-p)G_{in}(x)$ where $G_{in}(x)$ is the empirical distribution associated with the n random observations from the population π_i. The existence of some $\hat{\underset{\sim}{\alpha}}_i$ such that $S_i(\underset{\sim}{\alpha};p)$ attains the infimum can be shown by going through the analogous arguments as in [3]. Define $\hat{\underset{\sim}{\alpha}}_i$ to be such that

$$(3.2) \qquad S_i(\hat{\underset{\sim}{\alpha}}_i;p) = \inf_{\underset{\sim}{\alpha}\in\langle 0,1\rangle^m} S_i(\underset{\sim}{\alpha};p).$$

For a given value of p ($0 \leq p \leq 1$), we define a selection procedure R_p [illegible] follows.

[illegible] n independent observations from each π_i and construct the empirical distribution $G_{in}(x)$. Compute $\hat{\underset{\sim}{\alpha}}_i = \hat{\underset{\sim}{\alpha}}_i(X_{i1},X_{i2},\ldots,X_{in})$ which is defined by (3.1) and (3.2). Let $\lambda_{[1]}(\hat{\underset{\sim}{\alpha}}) \leq \lambda_{[2]}(\hat{\underset{\sim}{\alpha}}) \leq \ldots \leq \lambda_{[k]}(\hat{\underset{\sim}{\alpha}})$ denote the ordered values of $\lambda(\hat{\underset{\sim}{\alpha}}_1), \lambda(\hat{\underset{\sim}{\alpha}}_2),\ldots,\lambda(\hat{\underset{\sim}{\alpha}}_k)$.

R_p: Select π_i if, and only if $\lambda(\hat{\underset{\sim}{\alpha}}_i) \geq \lambda_{[k-t+1]}(\hat{\underset{\sim}{\alpha}})$

Use a mechanism when a tie occurs.

By a correct selection (CS) we mean a set of t populations asssociated with the t largest values of $\lambda(\hat{\underset{\sim}{\alpha}}_1), \lambda(\hat{\underset{\sim}{\alpha}}_2),\ldots,\lambda(\hat{\underset{\sim}{\alpha}}_k)$ is selected.

<u>**Definition 3.1**</u> A selection procedure R is consistent with respect to $(\mathfrak{F},\lambda)$ if $\lim_{\Delta\to 0}\lim_{n\to\infty}\inf_{\underset{\sim}{\alpha}\in\Omega(\lambda;\Delta)} P_{\underset{\sim}{\alpha}}\{CS|R\} = 1$

<u>**Definition 3.2**</u> A selection procedure R is strongly asymptotically monotone with respect to $(\mathfrak{F},\lambda)$ if $\lambda(\underset{\sim}{\alpha}_i) < \lambda(\underset{\sim}{\alpha}_j)$ and for any $\epsilon > 0$ implies

$$\lim_{n\to\infty} \sup_{\underset{\sim}{\alpha} \in \Omega(\lambda;\Delta)} P_{\underset{\sim}{\alpha}}\{\pi_i \text{ is selected } |R\} - \epsilon < \lim_{n\to\infty} \inf_{\underset{\sim}{\alpha} \in \Omega(\lambda;\Delta)} P_{\underset{\sim}{\alpha}}\{\pi_j \text{ is selected } |R\}.$$

<u>Theorem 3.1</u> For any value of p, $0 \le p \le 1$, R_p is consistent and strongly asymptotically monotone with respect to $(\mathfrak{F},\lambda)$.

Proof: (a) We show that for any p ($0 \le p \le 1$) and for each i $(i=1,2,\ldots,k)$, $\hat{\underset{\sim}{\alpha}}_i \to \underset{\sim}{\alpha}_i$ with probability one. We follow the arguments given in [3] with appropriate modifications. Now, by the Glivenko-Cantelli theorem, for $\epsilon > 0$, $\exists$ $N(\epsilon)$ such that, whenever $n \ge N(\epsilon)$,

$$P\{|p\ \underset{\sim}{\alpha}_i\cdot\underset{\sim}{F}+(1-p)G_{in}(x)-G_{in}(x)| < \epsilon\} = P\{p|\underset{\sim}{\alpha}_i\cdot\underset{\sim}{F} - G_{in}(x)| < \epsilon\} = 1.$$

Replacing $dF_n(x)$ by $d(p\underset{\sim}{\alpha}_i\cdot\underset{\sim}{F} + (1-p)G_{in}(x))$ and following the same argument as given in the proof of Theorem 2 in [3], the result follows.

(b) Consistency of R_p

Since λ is continuous it is is true that $\lambda(\hat{\underset{\sim}{\alpha}}_i) \to \lambda(\hat{\underset{\sim}{\alpha}}_i)$ <u>WP1</u> $(i=1,2,\ldots,k)$. Now, by the Egoroff's theorem, for $\epsilon > 0$ and $\delta > 0$, there exists $N_i(\epsilon,\delta)$, A_i and B_i such that the sample space is decomposed to be $A_i \cup B_i$ with B_i the complement of A_i and $P(B_i) > 1-\epsilon$ and on B_i, $|\lambda(\hat{\underset{\sim}{\alpha}}_i)-\lambda(\underset{\sim}{\alpha}_i)| < \delta$ whenever $n \ge N_i(\epsilon,\delta)$ uniformly in $\underset{\sim}{\alpha}_i \in \langle 0,1\rangle^m$, i.e, $N(\epsilon,\delta)$ is independent of $\underset{\sim}{\alpha}_i$. Note that $\lambda(\hat{\underset{\sim}{\alpha}}_i)$ depends on n. Set $N = N_1(\epsilon,\delta) +\ldots+ N_k(\epsilon,\delta)$ and set $B = \bigcap_{i=1}^{k} B_i$. Then, $P(B) > 1 - \epsilon$, and on B, whenever $n \ge N$, $\max_{1\le i\le k} |\lambda(\hat{\underset{\sim}{\alpha}}_i) - \lambda(\underset{\sim}{\alpha}_i)| < \delta$ uniformly for each $(\underset{\sim}{\alpha}_1,\underset{\sim}{\alpha}_2,\ldots,\underset{\sim}{\alpha}_k) \in \Omega$ (defined by (2.6)). Now, for any given $P^* \in (0,1)$, and for given $\Delta > 0$, however small, choose $\delta = \frac{\Delta}{3} > 0$ and $\epsilon = 1 - P^*$. Since on $\Omega(\lambda;\Delta)$, $\lambda_{[k-t+1]} - \lambda_{[k-t]} \ge \Delta = 3\delta$. Hence we conclude that

$$P_{\underset{\sim}{\alpha}}\{\lambda(\hat{\underset{\sim}{\alpha}}_{r_i}) > \lambda_{[k-t]}(\hat{\underset{\sim}{\alpha}}),\ i=1,2,\ldots,t|\lambda(\underset{\sim}{\alpha}_{r_i}) > \lambda_{[k-t]}(\underset{\sim}{\alpha})\} \ge P^*$$

for every $\underset{\sim}{\alpha} \in \Omega(\lambda;\Delta)$. Hence, we have shown that for every $\Delta > 0$, $\lim_{n\to\infty} \inf_{\underset{\sim}{\alpha} \in \Omega(\lambda;\Delta)} P_{\underset{\sim}{\alpha}}\{CS|R_p\} = 1$. Hence the consistency is shown.

(c) Suppose $\lambda(\underset{\sim}{\alpha}_i) < \lambda(\underset{\sim}{\alpha}_j)$.

(i) If $\lambda(\underset{\sim}{\alpha}_i) \leq \lambda_{[k-t]}(\underset{\sim}{\alpha})$ and $\lambda(\underset{\sim}{\alpha}_j) \geq \lambda_{[k-t+1]}(\underset{\sim}{\alpha})$. Then, take $P^* \geq 2/3$ and go through the arguments given in (b), we conclude that $\inf_{\underset{\sim}{\alpha} \in \Omega(\lambda;\Delta)} P_{\underset{\sim}{\alpha}}\{\pi_j$ is selected $|R_p\} \geq \inf_{\underset{\sim}{\alpha} \in \Omega(\lambda;\Delta)} P_{\underset{\sim}{\alpha}}\{CS|R_p\} \geq 2/3$ whenever $n \geq N_0 = N_0(\Delta)$ for some N_0. On the other hand, for each $n \geq N_0$, $\{\pi_i$ is selected $|R_p\}$ $\subset\{$Selection is not correct $|R_p\}$. Hence $P_{\underset{\sim}{\alpha}}\{\pi_i$ is selected $|R_p\} \leq 1 - P_{\underset{\sim}{\alpha}}\{CS|R_p\} \leq 1/3$ $\forall\ \underset{\sim}{\alpha} \in \Omega(\lambda;\Delta)$, i.e.

$$\sup_{\underset{\sim}{\alpha} \in \Omega(\lambda;\Delta)} P_{\underset{\sim}{\alpha}}\{\pi_i \text{ is selected } |R_p\} \leq 1/3 \text{ for each } n \geq N_0.$$

(ii) Suppose both $\lambda(\underset{\sim}{\alpha}_i)$ and $\lambda(\underset{\sim}{\alpha}_j)$ are no larger than $\lambda_{[k-t]}(\underset{\sim}{\alpha})$. Then, for $\epsilon > 0$ and by the arguments in (b), there exists a subset of sample space B and an integer N_0 such that $P\{B\} > 1 - \frac{\epsilon}{2}$ and for $n \geq N_0$ and on B, $\max_{1<i<k} \{|\underset{\sim}{\alpha}_i - \hat{\underset{\sim}{\alpha}}_i|\} < \frac{\Delta}{3}$. Let E denote the event $\{\pi_i$ is selected $|R_p\}$. Then $E = E \cap B + E \cap B^c$. Hence, $\sup_{\underset{\sim}{\alpha}} P_{\underset{\sim}{\alpha}}\{E\} \leq \sup_{\underset{\sim}{\alpha}} P_{\underset{\sim}{\alpha}}\{E \cap B\}$ $+ \sup_{\alpha} P_{\underset{\sim}{\alpha}}\{E \cap B^c\} \leq \sup_{\alpha} P_{\underset{\sim}{\alpha}}\{E \cap B\} + \frac{\epsilon}{2}$.
since $P_{\underset{\sim}{\alpha}}\{E \cap B^c\} \leq P_{\underset{\sim}{\alpha}}\{B^c\} < \frac{\epsilon}{2}$ $\forall\ \underset{\sim}{\alpha} \in \Omega(\lambda;\Delta)$. Noting that for any $\underset{\sim}{\alpha} \in \Omega(\lambda;\Delta)$, $P_{\underset{\sim}{\alpha}}\{E \cap B\} = 0$ since on B, $\hat{\underset{\sim}{\alpha}}_i < \underset{\sim}{\alpha}_{[k-t+1]} - \frac{\Delta}{3}$.

(iii) If $\lambda(\underset{\sim}{\alpha}_i)$ and $\lambda(\underset{\sim}{\alpha}_j)$ both are no less than $\lambda_{[k-t+1]}(\underset{\sim}{\alpha})$. The proof is analogous to the case of (ii).

The proof is thus complete.

Remark 3.1. If $t_1, t_2, \ldots, t_m$ are some positive integers such that each t_i is no larger than $k - 1$. Let $\Omega(t_1, t_2, \ldots, t_m) \equiv \{(\underset{\sim}{\alpha}_1, \underset{\sim}{\alpha}_2, \ldots, \underset{\sim}{\alpha}_k): \alpha^{(i)}_{[k-t_i+1]} > \alpha^{(i)}_{[k-t_i+1]}\ i = 1,2,\ldots,m\}$ where $\alpha^{(i)}_{[j]}$ denotes the j-th largest value of the i-th component of $\underset{\sim}{\alpha}_1, \underset{\sim}{\alpha}_2, \ldots, \underset{\sim}{\alpha}_k$ and we denote $\underset{\sim}{\alpha}_r = (\alpha^{(1)}_r, \alpha^{(2)}_r, \ldots, \alpha^{(m)}_r)$. If for each i we are desired to select the t_i largest in the i-th component simultaneously, then, using the statistics $\{\hat{\underset{\sim}{\alpha}}_1, \hat{\underset{\sim}{\alpha}}_2, \ldots, \hat{\underset{\sim}{\alpha}}_k\}$, which are defined by (3.2), associated with the i-th component, we select these populations which have the t_i largest values in the i-th component of $\{\alpha^{(i)}_1, \alpha^{(i)}_2, \ldots, \alpha^{(i)}_k\}$ $(i=1,2,\ldots,m)$. It can be shown that the simultaneous selections are also consistent and strongly asymptotically monotone on the configuration $\Omega(t_1, t_2, \ldots, t_k)$.

Remark 3.2. For $m = 2$ and a given n, let $\hat{\alpha}'_i$, $\hat{\alpha}''_i$ and $\hat{\alpha}_i$ denote respectively, the least square estimates associated with $p = 0$, $p = 1$ and some $p(0 < p < 1)$. Then, it can be obtained

$$\hat{\alpha}'_i = \Sigma(F_2-F_1)(F_1-\tfrac{i}{n})/\Sigma(F_2-F_1)^2,\quad \hat{\alpha}''_i = \hat{\alpha}'_i + \frac{1}{2n}\ \frac{\Sigma(F_2-F_1)}{\Sigma(F_2-F_1)^2}$$

and

$$\hat{\alpha}_i = \hat{\alpha}'_i + \frac{1-p}{2n}\ \frac{\Sigma(F_2-F_1)}{\Sigma(F_2-F_1)^2}$$

where

$$\Sigma(F_2-F_1) \equiv \sum_{i=1}^{n} (F_2(X_{[i]};\theta_2)-F_1(X_{[i]};\theta_1)) \text{ and } X_{[1]} \le X_{[2]} \le \cdots \le X_{[n]}$$

are the order statistics from π_i. As a convention we take $\hat{\alpha}'_i = 0$ if $\hat{\alpha}'_i < 0$ and $= 1$ if $\hat{\alpha}'_i > 1$ and use the same convention for other two cases. It can be seen that $\hat{\alpha}_i$ is always between $\hat{\alpha}'_i$ and $\hat{\alpha}''_i$ for all n. If F_1 and

F_2 are "smooth" in some sense, we see that $|\hat{\alpha}_i' - \hat{\alpha}_i''| = O(n^{-1+\epsilon})$ for $\epsilon > 0$.

<u>Definition 3.3</u> A selection procedure R is consistent of order $O(A(\Delta))$ $(o(A(\Delta)))$ with respect to $(\mathfrak{F},\lambda)$ if

$$\lim_{\substack{\Delta\to 0 \\ n=O(A(\Delta))}} \inf_{\underset{\sim}{\alpha}\in\Omega(\lambda;\Delta)} P_{\underset{\sim}{\alpha}}\{CS|R\} = 1 \qquad (\lim_{\substack{\Delta\to 0 \\ n=o(A(\Delta))}} \inf_{\underset{\sim}{\alpha}\in\Omega(\lambda;\Delta)} P_{\underset{\sim}{\alpha}}\{CS|R\} = 1).$$

<u>Theorem 3.2</u> For given p, $0 \le p \le 1$, R_p is consistent of order $O(\Delta^{\frac{-2}{1-2\delta}})$, $0 < \delta < 1/2$.

Proof: We note that, by the Glivenko-Cantelli theorem that $\sup_x |G_i(x)-G_{in}(x) + o(1)| \to 0$ WP1 as $n \to \infty$ $\forall_i$, where o(1) is independent of x. For any fixed i $(1 \le i \le k)$, let $\dot{S}(\underset{\sim}{\alpha}_i;p)$ denote the m-1 equations for which each equation is differentiated with respect to α_{ij}, $j = 1,2,\ldots$, m-1, where $\underset{\sim}{\alpha}_i = (\alpha_{i1}, \alpha_{i2},\ldots, \alpha_{im-1}, 1 - \sum_{j=1}^{m-1} \alpha_{ij})$. Then, the first element of $\dot{S}(\underset{\sim}{\alpha}_i;p)$ for j=1 becomes

$$\frac{1}{n}\sum_{j=1}^{n} F_1(X_{i[j]};\theta_1)\{\sum_{r=1}^{m}\alpha_{ir}F_r(X_{i[j]};\theta_r)-\frac{j}{n} + \frac{1-p}{2n}\}$$

$$\le \sup_x|G_i(X)-G_{im}(x)+o(1)|\frac{1}{n}\sum_{j=1}^{n} F_1(X_{i[j]};\theta_1).$$

where $X_{i[j]} \le \cdots \le X_{i[n]}$ are order statistics from π_i. Follow the analogous arguments of the proof of Theorem 4 of [3], we conclude that $|\hat{\underset{\sim}{\alpha}}_i - \underset{\sim}{\alpha}_i| < O(n^{-\frac{1}{2}+\delta})$ for all but finite n with prob. 1 where $0 \le \delta < \frac{1}{2}$. Now, if we take $\Delta/2 = O(n^{-\frac{1}{2}+\delta})$ for large n, we see that $n = O(\Delta^{\frac{-2}{1-2\delta}})$

and for this n it can be sure that the selection is correct with probability one as $\Delta \to 0$. The proof is thus complete.

Let $\bar{\underset{\sim}{\alpha}}_i$ denote the arithmetic mean of r independent estimates of $\hat{\underset{\sim}{\alpha}}_i$ where r is some integer. This means rn samples are drawn from each population. And for each subgroups of n samples, we obtain an estimate $\hat{\underset{\sim}{\alpha}}_i$ for the population π_i. If n is large, $\lambda(\underset{\sim}{\alpha}_i) = \alpha_{i1}$, and $t = 1$, we propose the following rule R_p'.

R_p': Select π_i if $\bar{\underset{\sim}{\alpha}}_{i1} \geq \bar{\underset{\sim}{\alpha}}_{j1}$ for all $j \neq i$.

where $\bar{\underset{\sim}{\alpha}}_{i1}$ is the first component of $\bar{\underset{\sim}{\alpha}}_i$.

<u>Theorem 3.3</u> If n is large, $t = 1$, and $\lambda(\underset{\sim}{\alpha}_i) = \alpha_{i1}$, the projection function, then we have

$$\inf_{\underset{\sim}{\alpha} \in \Omega(\lambda;\Delta)} P_{\underset{\sim}{\alpha}}\{CS|R_p'\} \geq \int_{-\infty}^{\infty} \prod_{j=2}^{k} \Phi(\delta_j z + \frac{\sqrt{r}\,\Delta}{\sigma_{[j]}})d\Phi(z)$$

where $\Phi(x)$ denotes the standard normal distribution and

$$\sigma_j^2 = 2 \int\int_{-\infty<x<y<\infty} G_j(x)[1-G_j(y)]dB_j(x)dB_j(y)$$

where

$$B_j(x) = F_1(x;\theta_1)G_j(x) - \int_{-\infty}^{\infty} F_1(x;\theta_1)dG_j(x)$$

for $j=1,2,\ldots,k$.

and $\quad \sigma_{[1]} \leq \sigma_{[2]} \leq \cdots \leq \sigma_{[k]}, \; \delta_j = \alpha_{[1]}/\sigma_{[j]}$.

Proof: It has been shown in [2] that $\hat{\underset{\sim}{\alpha}}_i$ is asymptotically normal and hence, the first component of $\hat{\underset{\sim}{\alpha}}_i$, say, $\hat{\alpha}_{i1}$ is asymptotically normal with mean α_{i1} and variance $\sigma_i^2 = 2 \int\int_{-\infty<x<y<\infty} G_i(x)[1-G_i(y)]dB_i(x)dB_i(y)$ where

$$B_i(x) = F_1(x;\theta_1)G_i(x) - \int_{-\infty}^{\infty} F_1(x;\theta_1)dG_i(x).$$

Hence, when n is large, t = 1, we have for $\underset{\sim}{\alpha}\in\Omega(\lambda;\Delta)$

$$P_{\underset{\sim}{\alpha}}\{CS|R_p'\} = P_{\underset{\sim}{\alpha}}\{\bar{\alpha}_{k1} \geq \bar{\alpha}_{j1}\ j=1,2,\ldots,k-1|\alpha_{k1} = \max_{1\leq j\leq k} \alpha_{j1}\}$$

$$= P_{\underset{\sim}{\alpha}}\{\frac{\sqrt{r}(\bar{\alpha}_{k1}-\alpha_{k1})}{\sigma_k} \geq \frac{\sqrt{r}(\bar{\alpha}_{j1}-\alpha_{j1})}{\sigma_j}\frac{\sigma_j}{\sigma_k} + \frac{\sqrt{r}(\alpha_{j1}-\alpha_{k1})}{\sigma_k}$$

$$\geq P_{\underset{\sim}{\alpha}}\{Z_k \geq Z_j(\frac{\sigma_j}{\sigma_k}) - \frac{\sqrt{r}\,\Delta}{\sigma_k}\quad j=1,2,\ldots,k\}$$ (where $Z_1,Z_2,\ldots,Z_k$ are iid standard normal)

$$= \int_{-\infty}^{\infty}\prod_{j=1}^{k-1}\Phi(\frac{\sigma_k}{\sigma_j}z + \frac{\sqrt{r}\,\Delta}{\sigma_j})d\Phi(z)$$

$$\geq \int_{-\infty}^{\infty}\prod_{j=1}^{k-1}\Phi(\delta_j z + \frac{\sqrt{r}\,\Delta}{\sigma_{[j+1]}})d\Phi(z)$$ (by a lemma in [4])

where $\delta_j = \sigma_{[1]}/\sigma_{[j+1]},\ \sigma_{[1]} \leq \sigma_{[2]} \leq \cdots \leq \sigma_{[k]}$.

This completes the proof.

Asymptotic relative efficiency of R_p with respect to a procedure R_B

We assume m=2, t=1, and λ is a projection function. In this case we have $G_i(x) = \alpha_i F_1(x;\theta_1) + (1-\alpha_i)F_2(x;\theta_2)$ for $i=1,2,\ldots,k$ and we denote α_i instead of $\underset{\sim}{\alpha}_i$. Suppose $F_1(x;\theta_1)$ and $F_2(x;\theta_2)$ are not specified, however, we assume there exists some point x_0, known, such that $F_1(x_o;\theta_1) \neq F_2(x;\theta_2)$. Assume $F_1(X_0;\theta_1) > F_2(X_0;\theta_2)$. Then, we see that $\alpha_i > \alpha_j$ if, and only if $G_i(x_o) > G_j(x_o)$. Hence, selecting best is equivalent to selecting the population associated with the largest $G(x_o;\alpha_i)$ value.

For a given i, $1 \leq i \leq k$, and j, $1 \leq j \leq n$, define

$$Y_{ij} = \begin{cases} 1 & \text{if } X_{ij} \leq X_o \\ 0 & \text{otherwise} \end{cases}$$

and define

$$\hat{G}_i(X_o) = \sum_{j=1}^{n} Y_{ij}.$$

Then, it is obvious that $\hat{G}_i(X_o)$ is binomial random variable with cdf $B(n;G(x_o))$.

We define a selection procedure R_B as follows:

R_B: Select the population π_i which is associated with the largest $\hat{G}_i(x_o)$.

When n is large, we use the normal approximation. Let $F_1(X_o;\theta_1) - F_2(X_o;\theta_2) = d_o > 0$. Then, by the result of [10], we have, asymptotically

$n \approx c^2(p^*)(1-\Delta^2 d_o^2)/2\Delta^2 d_o^2$, when $\Delta \to 0$, and $p^* \to 1$. Again, by the Feller's inequality, we see that $\Phi(z) \approx 1 - \frac{1}{\sqrt{2\pi}\, z} e^{-\frac{z^2}{2}}$. We obtain thus $C^2(p^*) = (\frac{1}{1-p^*})^2$. Let n_1 and n_2 denote, respectively, the sample sizes associated with R_p and R_B when $\inf_{\alpha\in\Omega(\lambda;\Delta)} p_{\underset{\sim}{\alpha}}\{CS\} \geq P^*$ is satisfied for both rules. We define the asymptotic relative efficiency of R_p with respect to R_B by $ARE(R_p;R_B) = \frac{n_1(p^*,\Delta)}{n_2(p^*,\Delta)}$ as $p^* \to 1$ and then $\Delta \to 0$. It follows from the previous result and Theorem 3.2. We have

$$ARE(R_p;R_B) = \lim_{\Delta\to 0} \lim_{P^*\to 1} \frac{C\Delta^{\frac{-2}{1-2\delta}}}{\frac{1}{2d_0^2\Delta^2(1-P^*)^2}} = 0.$$

However, if we take $1-P^* \equiv a = \Delta \to 0$, we have our another kind of effieiency given by

$$\text{ARE}(R_p;R_B) = \lim_{\Delta\to 0} \Delta^{2-\delta} = 0 \qquad \text{for } 0 < \delta < 1/2.$$

This shows that R_p is good compared to R_B. Also R_p holds for any general m and t. We should note that the case m = 2 and m > 2 are quite different and R_B is useful only for m = 2.

(B) Discrete Case

In this case, we denote $F_1, F_2,\ldots,F_m$ as discrete distribution such that the outcomes from each distribution with cdf F_i, for some i, can be classified into $s\,(\geq 2)$ states. Let the probability that an outcome from F_i belongs to state ℓ be denoted by $p_{i\ell}$. We assume $F_1, F_2,\ldots,F_m$ are all specified and $p_{i\ell}$ are all given.

For $\underset{\sim}{\alpha}_i \in \langle 0,1\rangle^m$ we define a mixture distribution G_i by

$$G_i(x) = \alpha_{i1}F_1(x) + \alpha_{i2}F_2(x) + \ldots + \alpha_{im}F_m(x).$$

Then, $G_i(x)$ is also a discrete distribution such that the probability of an outcome belonging to state j is given by

$$g_{ij} = \alpha_{i1}p_{1j} + \alpha_{i2}p_{2j} + \ldots + \alpha_{im}p_{mj}, \qquad \text{for } j=1,2,\ldots,s.$$

We assume there exists a lower bound g_o such that $g_{ij} \geq g_o > 0$ for all $i=1,2,\ldots,k$, $j=1,2,\ldots,s$. Let n samples be drawn from π_i and let n_j denote the number of outcomes which belong to state j. For any $\underset{\sim}{\alpha} = (\alpha_1,\ldots,\alpha_m)$, we define the Matusita distance (see [8]) as follows.

$$S_i(\underset{\sim}{\alpha}) = \{\sum_{j=1}^{s} (\sqrt{g_j} - \sqrt{\frac{n_j}{n}})^2\}^{\frac{1}{2}} \tag{3.3}$$

where $g_j = \sum_{i=1}^{m} \alpha_i p_{ij}$. $S_i(\underset{\sim}{\alpha})$ is thus a function on $\langle 0,1\rangle^m$.

Let $\hat{\underset{\sim}{\alpha}}_i$ denote a value in $<0,1>^m$ such that $S_i(\hat{\underset{\sim}{\alpha}}_i)$ attains its infimum, i.e. et $\hat{\underset{\sim}{\alpha}}_i$ be such that

$$(3.4) \qquad S_i(\hat{\underset{\sim}{\alpha}}_i) = \inf_{\underset{\sim}{\alpha} \in <0,1>^m} S_i(\underset{\sim}{\alpha}).$$

For given n and λ, to select the t best with respect to λ, we propose the following selection procedure.

R: Select $\pi_{r_1}, \pi_{r_2}, \ldots, \pi_{r_t}$ if, and only if,

$\lambda(\hat{\underset{\sim}{\alpha}}_{r_1}), \lambda(\hat{\underset{\sim}{\alpha}}_{r_2}), \ldots, \lambda(\hat{\underset{\sim}{\alpha}}_{r_t})$ are the t largest values of $\lambda(\hat{\underset{\sim}{\alpha}}_1)$, $\lambda(\hat{\underset{\sim}{\alpha}}_2), \ldots, \lambda(\hat{\underset{\sim}{\alpha}}_k)$ which are defined by (3.3) and (3.4). If there are ties, use a random mechanism.

<u>Theorem 3.4</u> The selection procedure R is consistent and strongly asymptotically monotone, with respect to $(\mathfrak{F}, \lambda)$.

Proof: It has been shown in Matusita [8] that for our case $\hat{\underset{\sim}{\alpha}}_i \to \underset{\sim}{\alpha}_i$ with probability one in the usual sense of convergence of a sequence of vectors. Therefore, $\lambda(\hat{\underset{\sim}{\alpha}}_i) \to \lambda(\underset{\sim}{\alpha}_i)$ WP1 for λ is continuous. Using the analogous arguments given in the proofs of Theorem 3.1, we can conclude the same results. This completes the proof.

REFERENCES

[1] Bartlett, M.S. and MacDonald, P.D.M. (1968) "Least-squares" estimation of distribution mixtures. Nature, Lond., 217 195-196.

[2] Choi, W. (1969) Estimates for the parameters of a finite mixture of distributions. Ann. Inst. Statist. Math. 21, 107-116.

[3] Choi, W. and Bulgren, W. G. (1968) An estimation procedure for mixtures of distributions. J. R. Statist. Soc. Ser. B 30, 444-460.

[4] Gupta, S. S. and Huang, W. T. (1974) A note on selecting a subset of normal populations with unequal sample sizes. Sankhyā, Ser. A, 36, 389-396.

[5] Hill, B. M. (1963) Information for estimating the proportions in mixtures of exponential and normal distributions. J. Amer. Statist. Assoc. 58, 918-932.

[6] Lehmann, E. L. (1951) Consistency and unbiasedness of certain non-parametric tests. Ann. Math. Statist. 22, 165-179.

[7] Macdonald, P.D.M. (1971) Comment on "An estimation procedure for mixtures of distributions" by Choi and Bulgren. J..R. Statist. Soc. Ser. B, 33, 326-329.

[8] Matusita, K. (1954) On the estimation by the minimum distance method. Ann. Inst. Statist. Math. 5, 59-65.

[9] Robbins, H. (1964) The empirical Bayes approach to statistical decision problems. Ann. Math. Statist. 35, 1-20.

[10] Sobel, M. and Huyett, M. (1957). Selecting the best one of several binomial populations. Bell System. Tech. J. 36, 537-576.

[11] Teicher, H. (1963) Identifiability of finite mixtures. Ann. Math. Statist., 34, 1265-1269.

[12] Wolfowitz, J. (1954). Estimation by the minimum distance method. Ann. Inst. of Statist. Math. 9-23.

[13] Yakowitz, S. and Spragins, J. (1968) On the identifiability of finite mixtures. Ann. Math. Statist., 39, 209-214.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: Mimeograph Series #504
2. GOVT ACCESSION NO.:
3. RECIPIENT'S CATALOG NUMBER:
4. TITLE (and Subtitle): On Selection Rules for Finite Mixtures of Distributions
5. TYPE OF REPORT & PERIOD COVERED: Technical
6. PERFORMING ORG. REPORT NUMBER: Mimeo. Series #504
7. AUTHOR(s): Shanti S. Gupta and Wen-Tao Huang
8. CONTRACT OR GRANT NUMBER(s): ONR N00014-75C-0455
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Purdue University, Department of Statistics, West Lafayette, IN 47907
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS:
11. CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research, Washington, DC
12. REPORT DATE: August 1977
13. NUMBER OF PAGES: 17
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office):
15. SECURITY CLASS. (of this report): Unclassified
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:
16. DISTRIBUTION STATEMENT (of this Report): Approved for public release, distribution unlimited.
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report):
18. SUPPLEMENTARY NOTES:
19. KEY WORDS (Continue on reverse side if necessary and identify by block number): Finite mixture of distributions, consistent selection procedure, strong asymptotic monotonicity, minimum distance, least squares.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A class of consistent selection procedures is proposed for the selection of the t largest proportions of finite mixture of distributions. The approach is based on the indifference zone formulation. Some asymptotic optimal properties are shown to hold. For special case of $m = 2$, the number of components in each of k mixture populations, some results of the asymptotic relative efficiency are obtained.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)